International Journal of Engineering Research and General Science Volume 3, Issue 3, May-June, 2015
ISSN 2091-2730

# Precognition of Students Academic Failure Using Data Mining Techniques

Sadaf Fatima Salim Attar, Prof. Y.C Kulkarni

Information Technology, Bharati Vidyapeeth Deemed University College of Engineering, Pune, India

sadaf17attar@gmail.com Phone No. +918237260070

**Abstract**—This paper proposes to pre-recognize student's academic failure. Real time data on school or graduating students from an institute is taken and various data mining techniques (classification algorithms), such as induction rules, decision trees and naive bayes are applied on it. The results of these algorithms are being compared and optimized for foretelling which students might fail in future. We first consider all the available attributes of students, then select few best attributes and finally, rebalance the data using classification algorithms. The use of data mining concept in the field of education is called as Educational Data Mining, EDM [2]. This paper focuses on designing various methods that will help the teachers and the principal (Administrator) of the school to figure out the weak students and improve their educational standards and environment in which they learn. I propose the use of data mining procedures, because the complexity of the problem is high, data to be handled is very large and often highly unbalanced. The final objective of this paper is to detect the failure of students as early as possible to prevent them from dropping out and improve their academic performance. The outcomes are compared and the best results are shown.

**Keywords**— Data Mining, Educational Data Mining, Decision Trees, Induction Rules, Rebalancing Data, Classification Algorithms.

## INTRODUCTION

Many educational organizations and school administrations today, leave no stone unturned to improve their student's academic performance. They want to increase the number of student's getting passed in their yearly academics. The reason for this is to maintain the brand name of the organization and as well as to educate students in a better way. In order to increase the number of students getting passed, we have to first find out the students that may get failed in that particular year in academics. This project basically aims to foretell the student's failure beforehand, so that some measures can be taken to avoid the student's failure in future.

To predict the failure of students is a complex task, as it requires large number of the data to be handled. We need to maintain the record of students each and every activities that he/she does in his/her day to day life. Based on this information, and applying some data mining algorithms on it, we may be able to predict the student's failure.

Data mining is the abstraction of needful data from large databases and ignoring the rest. Data mining tools predict future trends and behaviors, allowing the organizations to make proactive, knowledge-driven decisions [3]. Data mining helps the people to make quick decisions on a situation as compared to statistical analysis. Data mining tools can easily handle large amount of data stored in datasets, they can pre-process the data, and can work on unbalanced data easily. Data mining basically uses more direct approach and does meta-heuristics search on data.

The scope of data mining is subjected to automated prediction of trends and behaviors. Artificial neural networks, decision trees, genetic algorithms, nearest neighbor method and induction rules are some of the most widely used methods of data mining [3]. This project makes use of two rules of induction, two decision tree algorithms of data mining and naive bayes algorithm (which is also a classification algorithm used for prediction). Data mining techniques have been under development for decades and are of huge use in research areas like statistics, artificial intelligence and machine learning [3].

This study proposes to foretell the student's academic failure using the algorithms of data mining techniques. The algorithms are applied on huge collection of data on student's activities and the results are obtained, through which the failure can be predicted. This information is more useful for the teachers and principal of the organization, so that they can make proper arrangements and facilities to increase the capability of students and reduce/prevent the failure of students in academics years. These experiments have shown almost expected results in context with economic, educational or sociological characteristics that may be helpful in foretelling low academic performance.

# I. METHODOLOGY

The main goal of my paper is to foretell the student's academic performance. This projects aims to pre-recognize the student's academic failure using data mining technique. The students may belong to any educational organization like higher secondary school or Graduation College.

In my paper some of DM algorithms will be used to foretell the student's failure so that proper attention can be given to those students who may fail in future. This project will help the instructors as well as students to improve their performance by adapting certain changes in the standards of their teaching methodologies.

Educational data mining is basically used which focuses on development of methods to better understand students and the environment in which they learn [8]. I am going to implement two rules of induction, two decision tree algorithms and naive bayes algorithm to predict the failure of student's.

For this project, I am going to implement the spiral model of the software development models. Spiral model is a combination of prototyping model and waterfall model [5]. This model is basically used for large projects and the projects that require continuous up gradation. The spiral model consists of four phases named as planning phase, risk analysis phase, development and testing phase, and evaluation phase. One iteration (spiral) consists of these four activities and the output of this is a small prototype of the large software. This prototype is checked to see if it meets the required expectations and then all the four activities are again repeated for all the spirals until the entire project is built.

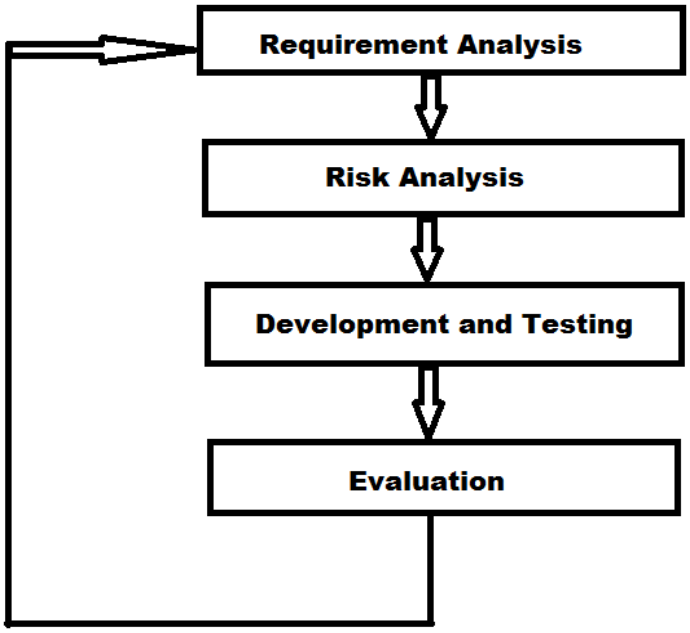


**fig. 1 SDLC Spiral Model**

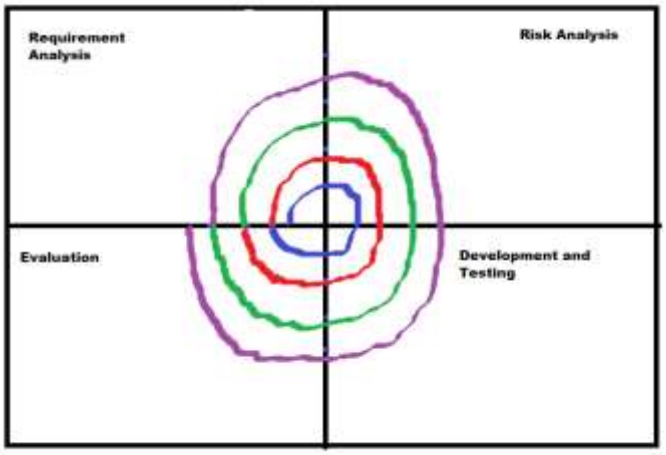


fig. 2 Spiral Model Design

In the (fig. 2) each iteration is represented by different color. The first iteration (spiral) is shown by blue color which covers all the four phases of spiral model (Requirement Analysis, Risk Analysis, Development and Testing, Evaluation). Once the evaluation phase for the first iteration (spiral) is completed, the second iteration (spiral) is started which is represented by red color, here again from requirement analysis to evaluation phase and so on until the entire project is build. The advantage of using spiral model is that development of the project is fast, risk factor is evaluated, customer feedback is taken and changes are implemented faster and so on. The disadvantage is that it is not suitable for smaller projects; spiral may go infinitely [5].

Now we will see modules of the project.

**1.1 Modules of the Project:**

There are four main modules of the project. They are as follows:

1. Data Collection
2. Data Management
3. Data Mining
4. Implementation

Data Collection is a process where information about the students is collected. This information is nothing but the data that will be useful in predicting the failure of students in academics. The data about students is collected in three different categories; first category is specific survey where personal and family information of the student is collected.
For example, number of hours spent studying daily, number of students in each batch, attendance of students in morning/evening tutorials, occupation of father and mother, number of members in a family, studying habits, any illness, etc. second category of data collection is academic information of the students. This data is the information that is required by various higher and secondary education institutions while admitting the students in their institutions. For example, age, gender, previous school information, type of school, marks in math, marks in English, marks in chemistry etc. The third category of data collection is departmental survey where each subject's department wise information of a student is collected. For example marks in math 1, marks in math 2, marks in English 1, and marks in English 2 etc. All this information is then stored in the dataset.

Data Management refers to preparing the data for applying data mining techniques. In data management, we do data cleansing, transformation of variables and data partitioning.
One of the most important techniques of data management is the selection of features (attributes) by applying feature selection algorithms. The attribute selection algorithm tries to select those features of students which have greater impact on their academic status. Few attribute selection algorithms are as follows, CfsSubsetEval, Filtered-AttributeEval, FilteredSubsetEval, etc. Because of these attribute selection algorithms we can select the best attributes out of huge number of attributes of students that affect the student's performance.

Data Mining consists of certain DM algorithms that help in predicting the student's failure using classification algorithms. The classification algorithms that we are going to use are two rules of induction algorithms; NNge (it is a nearest neighbor approach); OneR [1], which uses the minimum-error attribute for class prediction; and two decision tree rules; RandomTree [1], which considers K randomly chosen attributes at each node of the tree; SimpleCart [6], which implements minimal cost-complexity pruning. I am also using another classification algorithm called Naive Bayes Algorithm [7] provided by Microsoft SQL Server Analysis Services. This

algorithm is basically used for predictive modeling which is based on Bayesian Techniques. Finally, the results of all these executed algorithms evaluated, compared and optimized to determine which one gives the best result.

Implementation is the last phase of the project where the results obtained from DM techniques are interpreted into a model. For implementation, i am going to make use of .Net Technology.

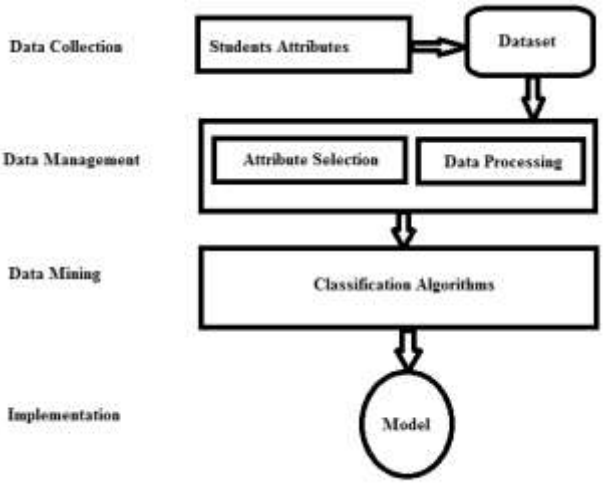


Fig. 3 Modules of the Project

### 1.2 Components of the Project

The components of the project are mainly divided into two parts: Functional Components and Non- Functional Components.

Functional Components are those components of the project whose actions/ results can be seen on screen. These are the entities whose actions can predict the failure of students in future. They are as follows:

1. Student
2. Teacher
3. Administrator
4. Prediction Tool

Student is the basic component of the project. The project mainly focuses on pre-recognizing the academic failure of the students so that proper guidance can be provided to those students who may fail in future and help them from dropping out. Each student registers itself on the site, and can fill its information. The information can be his/her personal information, academic information and department wise information. The students only have the authority to see their results and notices arranged for them by their teacher.

Teacher has a very important role in this project. The teacher is the only person who has access to the prediction tool. The teacher can view the results calculated by the prediction tool and take appropriate decisions regarding that particular student. The teacher can view the details of all the students, manage the lecture batches of the students, add/update other skill-sets of students, short-list the students, arrange exam schedule for students, arrange notices for the students and prepare a report.

Administrator can view the final result and arrange the notices.

Prediction Tool is a tool that calculates the number of students that may fail in future. The tool is basically based on data mining concept and consists of classification algorithm that calculates the failure of students. The classification algorithm is composed of two rules of induction algorithm, two decision tree algorithms and naive bayes Algorithm. The induction rule algorithms are NNge (it is a nearest neighbor approach) and OneR [1], which uses minimum-error attribute for class prediction and the decision tree algorithms are SimpleCart [6], which implements minimal cost-complexity feature and RandomTree [1], which considers K randomly chosen attributes at each node of the tree. Naive Bayes classification algorithm calculates the probability of every state of each input column, given each possible state of the predictable column [7]. The decision tree algorithms, induction rules and naive bayes algorithms can be easily implemented in the form of IF-THEN rules of object-oriented programming, which can be easily understood. In this way,

even a normal user who doesn't have any deep knowledge about data mining, for e.g. teacher and administrator can easily understand the results obtained using these algorithms.

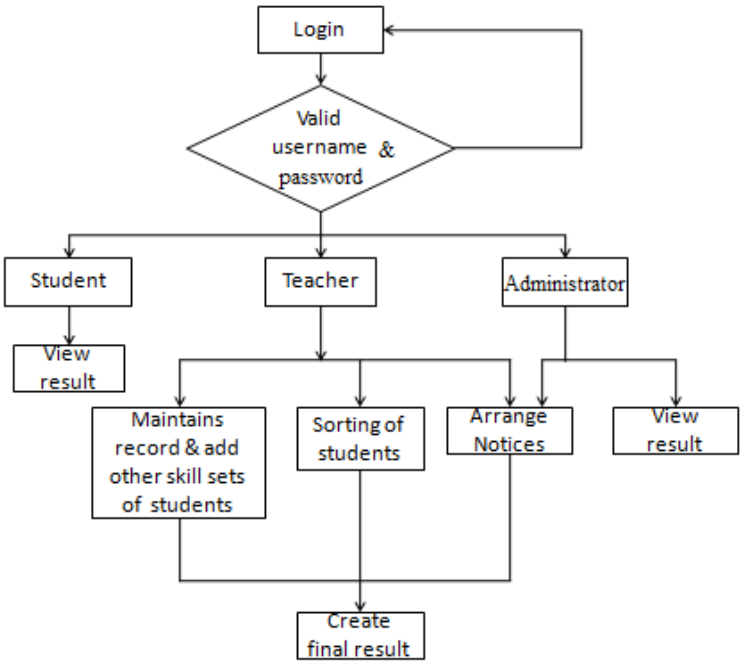


**Fig.4 Flow of the System**

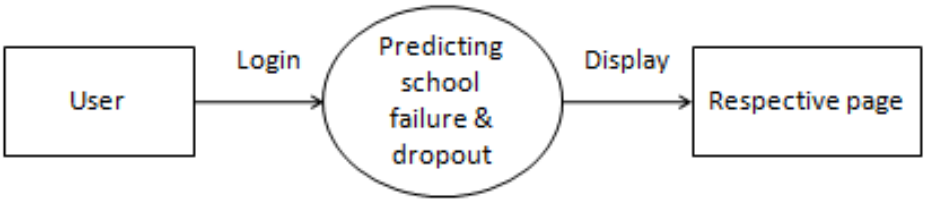


**Fig. 5 Level 0 DFD**

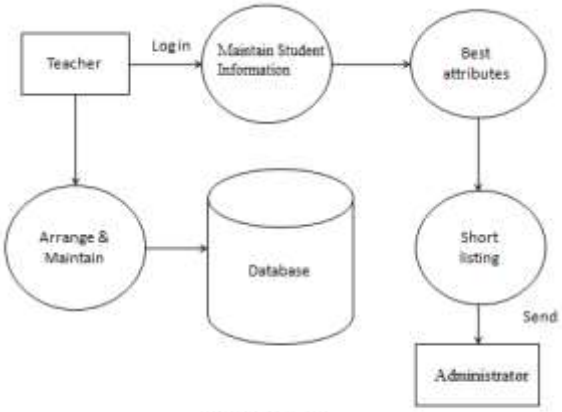


Fig. 6 Level 1 DFD

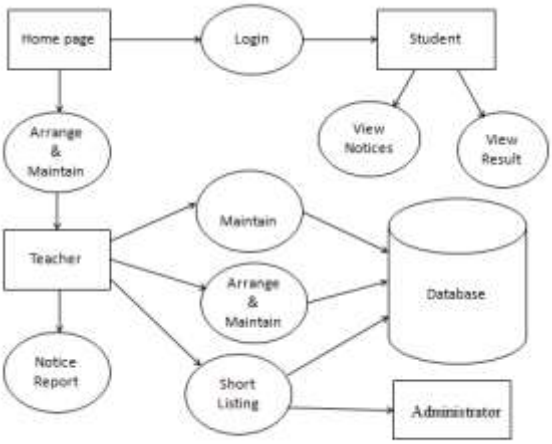


Fig. 7 Level 2 DFD

Non-Functional Components are those components of the project that run in background and whose actions can't be seen on screen. The non-functional components support the functional components and together they produce the final result of student's failure report.
They are as follows:

1. Data Collection Techniques
2. Attribute Selection
3. Dataset Management

Data Collection Techniques basically deal with gathering student related information that will be useful in predicting their failure in futute. The information is provided by the student itself.There are three categories in which this data is collected. First one is the specific survey where personal and family information of the student is collected. For example, number of hours spent studying daily, number of students in each batch, attendance during morning/evening tutorials, occupation of father and mother, number of members in a family, studying habits, any illness, etc. second category of data collection is academic information of the students. This information is the data that is required by various higher and secondary education institutions while admitting the students in their institutions. For example, age, gender, previous school information, type of school, marks in math, marks in English, marks in chemistry etc. The third category of data collection is departmental survey where each subject's department wise information of a student is collected. For example marks in math 1, marks in math 2, marks in English 1, and marks in English 2 etc. All this information is then stored in the dataset.

Attribute Selection basically deals with selecting the best attributes out of huge collection of attributes, based on which the results can be calculated. Practically, the information provided by each student is more than sufficient for the prediction. Instead of making use of this whole information for prediction, we can select few best attributes out of the huge collection and precede the further process of

prediction. This simplifies the complexity of the programmer and also the program. There will not be much difference in the results obtained. This step of attribute selection is only to ease the functionality.

Dataset Management deals with the management of data that is stored in the dataset. The information provided by the students may not be accurate or may not be precise. Dataset Management involves Data Cleaning, Integration and Discretization, and Variable Transformation. It also involves data redundancy, spelling mistakes, invalid data, etc. For example. "N" is to be transformed into "N ". Also in case where Age of student's should be set in the dd/mm/yy format. Another case is that numerical values of the marks obtained by students in each subject are to be changed to categorical values [1]. For e.g. for excellent scoring: score should be between 9.5 and 10, very good scoring: score should be between 8.5 and 9.4 and so on. And at last all the cleaned data is to be integrated into a dataset.

## CONCLUSION

Prior work on predicting student's academic failure was based on Weka tool. All the algorithms required for obtaining results were just outsourced by the previous system. Also the existing system implement five rules of induction and five decision tree algorithms which increased the complexity and overhead of the system. In this paper, we implemented the algorithms in the system on our own. We did not outsource the algorithms from Weka tool. Also we implemented only two rules of induction, two decision tree algorithms and naive bayes algorithm which decreased the complexity and overhead of the system. The selection of the features attributes of the student can be done manually or automatically using algorithms. We made this project a real-time application which can be used in any educational organization for pre-recognizing the failure of students.

## REFERENCES:

[1] Carlos Marquez-Vera, Cristobal Romero Morales, and Sebastian Ventura Soto, "Predicting school failure and dropout by using data mining techniques". IEEE Journal of Latin-American Learning Technologies, Vol. 8, No. 1, February 2013.

[2] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert Syst. Appl., vol. 33, no. 1, pp. 135-146, 2007.

[3]http://www.thearling.com/text/dmwhite/dmwhite.htm

[4] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

[5] http://www.softwaretestinghelp.com/spiral-model-what-is-sdlc-spiral-model/

[6] L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. New York , USA: Chapman & Hall, 1984.

[7] https://msdn.microsoft.com/en-us/library/ms174806.aspx

[8] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Trans.Syst., Man, Cybern. C, Appl. Rev.*, vol. 40, no. 6, pp. 601-618, Nov. 2010.

[9] Dr. Vuda Sreenivasarao, Capt. Genetu Yohannes, " Improving Academic Performance of Students of Defence University Based on Data Warehousing and Data Mining," Global Journal of Computer Science and Technology, Vol 12, Issue 2, Version 1.0, January 2012.

[10] M. N. Quadril and N. V. Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques," *Global J. Comput. Sci. Technol.*, vol. 10, pp.2-5, Feb.2010.

[11] A. Parker, "A study of variables that predict dropout from distance education," *Int. J. Educ. Technol*., vol. 1, no. 2, pp. 1-11, 1999.

[12] A Machine Learning Algorithms in Java, Written I, Frank E. WEKA, Morgan Kaufmann Publishers, 2000.

[13] Y. Freund and L. Mason, "The alternating decision tree algorithms," in Proc. 16th Int. Conf. Mach. Learn., 1999, pp. 124-133